The Way to my Heart is through Contrastive Learning: Remote Photoplethysmography from Unlabelled Video

John Gideon[*] Simon Stent[*]
Toyota Research Institute
Cambridge, MA, USA
{john.gideon, simon.stent}@tri.global

ENCODER	in	out	kernel	stride	pad
Conv3d + BN3d + ELU	C	32	(1,5,5)	1	(0,2,2)
AvgPool3d			(1,2,2)	(1,2,2)	0
Conv3d + BN3d + ELU	32	64	(3,3,3)	1	(1,1,1)
Conv3d + BN3d + ELU	64	64	(3,3,3)	1	(1,1,1)
AvgPool3d			(2,2,2)	(2,2,2)	0
Conv3d + BN3d + ELU	64	64	(3,3,3)	1	(1,1,1)
Conv3d + BN3d + ELU	64	64	(3,3,3)	1	(1,1,1)
AvgPool3d			(2,2,2)	(2,2,2)	0
Conv3d + BN3d + ELU	64	64	(3,3,3)	1	(1,1,1)
Conv3d + BN3d + ELU	64	64	(3,3,3)	1	(1,1,1)
AvgPool3d			(1,2,2)	(1,2,2)	0
Conv3d + BN3d + ELU	64	64	(3,3,3)	1	(1,1,1)
Conv3d + BN3d + ELU	64	64	(3,3,3)	1	(1,1,1)
DECODER					
Interpolate			(2,1,1)		
Conv3d + BN3d + ELU	64	64	(3,1,1)	1	(1,0,0)
Interpolate			(2,1,1)		
Conv3d + BN3d + ELU	64	64	(3,1,1)	1	(1,0,0)
AdaptiveAvgPool3d			(-,1,1)		
Conv3d	64	1	(1,1,1)	1	(0,0,0)

Table 1. **Modified PhysNet-3DCNN architecture**. The architecture follows an encoder-decoder structure with 3D convolutions to represent patterns through time; "s" corresponds to stride, "p" to padding, "C" to the number of input channels.

SALIENCY NET	in	out	kernel	stride	pad
Conv2d + BN2d + ReLU	C	64	(1,7,7)	2	3
MaxPool			(1,3,3)	2	1
BasicBlock	64	64		1	
BasicBlock	64	64		1	
BasicBlock	64	64		1	

Table 2. **Saliency Network**. The architecture follows a Resnet-18 structure, truncated after `layer1`, with pre-trained ImageNet weights. Each BasicBlock consists of a 3 × 3 convolution, 2D batch normalization, ReLU, 3 × 3 convolution, 2D batch normalization, addition with the BasicBlock input (the residual) and a final ReLU. For further details see [1].

A. Appendix

A1. Model Architecture

For the PPG estimator we use a modified 3D-CNN version of PhysNet [5] as described in Table 1. Our modification is to use interpolation and convolution in the decoder instead of transposed convolution, which we found to reduce the aliasing artifacts that were present in the original model. For the saliency sampler, we use the architecture described in [2], swapping out the saliency network for the shallower model shown in Table 2 which was found to be sufficient for detecting facial parts (by the nature of the saliency maps learned).

A2. Other Loss Functions/Metrics

Pearson's correlation (PC) is commonly used as a loss and metric in other rPPG works (*e.g.* [5]). While it is scale invariant, it assumes that there is perfect temporal synchronization between the ground truth and observed data. Otherwise the network must be capable of learning a temporal offset, assuming the offset is constant.

Signal-to-noise ratio (SNR) is another baseline used in prior rPPG works (*e.g.* [3]) which train to match a ground truth heart rate instead of the full PPG signal. It relaxes the alignment assumption by expressing the loss in the frequency domain using the power spectral density (PSD). It calculates the amount of power in the target heart rate frequency bin of the PSD and compares it to the total other power in the PPG signal. Because of this, it assumes that all other frequencies should be zeroed out, which may remove meaningful harmonics.

A3. Dataset PPG Performance

To supplement Table 3 from the main paper, we present further results on the four PPG datasets in Table 3, using

[*] Equal contribution

	PURE			COHFACE			MR-NIRP-Car			UBFC		
Method	PC	MCC	SNR	PC	MCC	SNR	PC	MCC	SNR	PC	MCC	SNR
Mean	-0.01	0.12	-9.6	-0.01	0.14	-9.2	-0.02	0.36	5.6	0.00	0.10	-13.0
Median	0.00	0.08	-10.9	0.00	0.14	-9.2	0.00	0.30	1.4	0.01	0.10	-13.5
Our Supervised	0.54	**0.90**	18.1	0.23	0.57	15.4	**0.52**	**0.79**	17.8	**0.17**	**0.64**	13.3
With Saliency	**0.58**	**0.90**	18.1	**0.30**	0.57	15.5	0.42	0.78	17.9	0.15	**0.64**	13.3
Our Contrastive	0.02	0.79	19.4	-0.19	**0.65**	17.9	0.18	0.74	**19.3**	0.03	0.63	**14.3**
With Saliency	0.00	0.80	**19.5**	-0.04	**0.65**	**17.9**	0.27	0.74	18.8	0.01	0.61	13.3

Table 3. **Experiment PPG Statistics.** PPG statistics on all datasets using our supervised and contrastive systems, with and without saliency. We compare with both a mean and median baseline. The top performing system varies greatly depending on the dataset and statistic. However, the contrastive systems often perform comparable or better than the supervised ones when considering sync-robust metrics (without the need for ground truth).

metrics which capture statistics of predicted vs. ground truth PPG signals (as opposed to the final predicted heart rate). We again show the results of both our supervised and contrastive systems, with and without the use of a saliency sampler. However, as these PPG statistics are not given for the other cited baseline systems, we instead provide only the mean and median methods as baselines.

We calculate the Pearson's correlation (PC), Max cross-correlation (MCC), and Signal-to-noise ratio (SNR) of the predicted versus ground truth PPG signals. While we present PC for comparison, we expect MCC and SNR to be better measures of system performance, as they are calculated in the frequency domain. This makes them more robust to desynchronization between predictions and ground truth. Supervised training uses MCC as a loss function and validation metric, while contrastive training works without guiding ground truth. Because of this, it is possible for either system to learn a random phase offset, as long as the overall frequency information is predictive of heart rate.

PC. Across all dataset results, we note that supervised training attains the highest PC. Even though desynchronization is not penalized during training, the model likely learns the easiest mapping between video and ground truth - one without an additional offset. This could indicate that the datasets only have minimal offset between observation and ground truth.

MCC. We also note that supervised training tends to result in higher MCC, which is likely due to the guiding ground truth. Without ground truth, the contrastive method is only able to learn the periodic signal visible in the input video. The supervised method would be able to learn to replicate any repeating artifacts of the biometric PPG sensor, producing a stronger MCC. However, the overall performance on the contrastive COHFACE model is substantially better than that of the supervised one, as seen in the heart rate results. This likely lessens the relative impact of PPG artifacts, when compared with other datasets with closer performance.

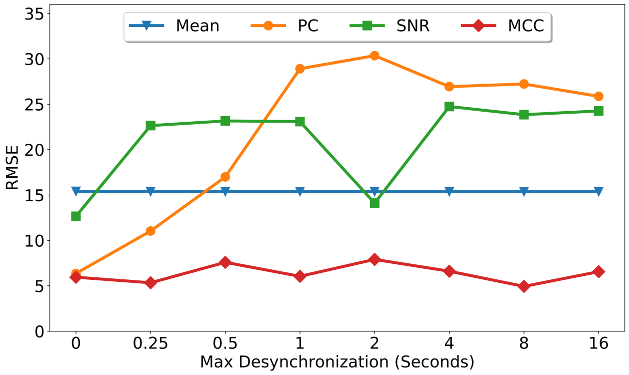

Figure 1. **Loss Function Robustness to Desynchronized Ground Truth.** The performance of supervised training on COHFACE with varying amounts of random desynchronization applied between the video and the ground truth PPG signal. We show the performance of 3 loss functions (PC, SNR, MCC) against a Mean model baseline (which estimates the mean heart rate in the test set). Unlike other loss functions, MCC is shown to be robust to ground truth desynchronization.

SNR. Unlike MCC, SNR penalizes the learning of all other frequencies besides heart rate. So perfect PPG prediction can result in lower SNR performance, if the PPG includes other frequencies. Because the supervised training uses MCC as a loss function and the ground truth isn't a perfect sine wave, this encourages sub-optimal SNR. We see this reflected in the results, with contrastive learning having a higher SNR across all datasets.

A4. Loss Function Robustness

In Section A3, we examined how the metrics PC, MCC, and SNR can be used to gauge the performance of PPG prediction versus ground truth. In this section, we examine how each metric performs as a loss function during supervised training, as well as the impact of desynchronization between the observed video and the ground truth PPG.

We use the COHFACE dataset and compare versus a baseline that always predicts the mean heart rate in the test set. We select the maximum amount of injected synchronization error (O_{max}) to range between 0 and 16 seconds. Each time a new clip is drawn during training, a random offset is chosen between $-O_{\text{max}}$ and O_{max} using uniform sampling. We then shift the ground truth PPG by the selected offset using neighboring data. We train a model using the supervised pipeline and the selected loss function. We calculate the selected loss on held-out validation data each epoch and use the model with the lowest loss at test time. In this experiment, we do not use the saliency sampler since the purpose is to explore the robustness of supervised loss functions.

Figure 1 shows the RMSE performance of our supervised system for different O_{max} and loss functions. Without injected desynchronization, we find that PC and MCC perform similarly. This likely indicates that ground truth in COHFACE is consistently aligned with the video, either with a minimal or constant (learnable) offset. Because MCC is the offset-adjusted version of PC, we note that they have similar performance without the presence of offsets. However, when increasing amounts of synchronization error are applied, the performance of the system trained with PC quickly degrades, while the one trained with MCC remains relatively stable. We also note that SNR has consistently poor performance at all offsets, indicating that it is a weaker supervisory signal for rPPG compared to correlation-based measures. Based on these results, we favored MCC as a loss function for supervised training, particularly if synchronization issues were suspected to be present in the training data.

A5. Dataset Statistics

Our approach relies on the assumption (Assumption 2 in the main paper) that “the signal of interest typically does not vary rapidly over short time intervals: the heart rate of a person at time t is similar to their heart rate at $t + W$, where W is in the order of seconds.” In Fig. 2, we show the distribution of heart rate variation at different time intervals within each of the four datasets. With approximately 80%-100% certainty, depending on the dataset, one can assume that the heart rate at $t + 10$s is within 10bpm of the heart rate at time t. All datasets consistently have a median variation of about 2.5bpm over a 10s period. UBFC was found to contain the most heart rate variability of the datasets examined.

A6. Unsupervised Learning Protocol

In our contrastive experiments, since our method does not utilize ground truth labels, we fold validation data into the training set. In other words, in Table 3 of the main paper, the contrastive models “see” a little more training data than the supervised models. When validation data is excluded from contrastive training, we found test RMSE to worsen by 0.7 bpm on average across datasets. This shows the value of larger training sets when training with contrastive loss. Exploring the trade off between training data size and model performance is a topic for future work.

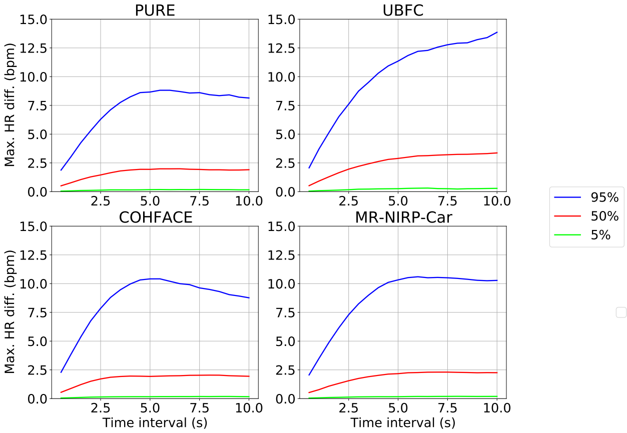

Figure 2. **Distribution of heart rate differences at different time intervals apart**. The 5%, 50% and 95% quantile lines are shown.

A7. Additional Baseline

While we selected the strongest comparable baseline we could find for each dataset, one reviewer requested the addition of the Siamese CNN proposed by [4]. However, we note that this work is not directly comparable for several reasons, in particular: (i) for UBFC evaluation, we do not pre-train our model on COHFACE; (ii) the final results reported in [4] use a 20s (COHFACE) and 30s (UBFC, PURE) input time window, while we use a shorter 10s window consistently across datasets. In Table 2 of [4] the effect of window length is shown for COHFACE: they report 1.8 RMSE with window length 400, and 4.7 with 256. Our *self-supervised* baseline achieved 4.6 RMSE with window length 300 (25 run average). This shows that over a range of datasets we can achieve performance comparable to supervised deep learning methods (*e.g.* [4]) using our approach *without annotations*.

A8. Sensitivity to Regularization Parameters

In Table 4 we show the relative performance of a contrastive model trained on the UBFC dataset as the saliency sparsity regularization weight, w_s, and temporal regularization weight, w_t, are varied in the range [0, 0.1, 1, 10]. We find that model performance is not significantly impacted. The saliency map output is most visible when higher values of the sparsity term w_s are used, although we observe that the best parameters can depend on variables such as dataset and resolution.

	RMSE change (↓)				MAE change (↓)				PC change (↑)			
$w_s \setminus w_t$	0	0.1	1	10	0	0.1	1	10	0	0.1	1	10
0	-	1.4	1.4	-0.9	-	0.7	0.6	-0.2	-	-0.06	-0.06	0.02
0.1	3.1	0.8	-0.8	0.1	1.3	0.4	-0.2	0.0	-0.18	-0.04	0.02	0.00
1	-0.8	0.0	-0.7	-0.4	-0.2	0.0	-0.2	-0.1	0.02	0.00	0.02	0.01
10	0.2	0.6	0.4	-0.9	0.0	0.2	0.2	-0.2	0.00	-0.02	-0.01	0.03

Table 4. **Sensitivity of saliency sampler regularization on rPPG performance for a contrastively trained model on UBFC.** As with Table 3 from the main paper, we average the results of five runs on different folds with held-out subjects, and report the performance differences relative to the zero regularization case, (w_s, w_t) = (0,0). rPPG performance tends to improve when the regularization parameters are set in the range [1,10], although the best parameters depend on variables such as dataset and resolution.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[2] Adria Recasens, Petr Kellnhofer, Simon Stent, Wojciech Matusik, and Antonio Torralba. Learning to zoom: a saliency-based sampling layer for neural networks. In *ECCV*, pages 51–66, 2018. 1

[3] Radim Špetlík, Vojtěch Franc, Jan Čech, and Jiří Matas. Visual heart rate estimation with convolutional neural network. In *BMVC*, 2018. 1

[4] Yun-Yun Tsou, Yi-An Lee, Chiou-Ting Hsu, and Shang-Hung Chang. Siamese-rPPG network: Remote photoplethysmography signal estimation from face videos. In *Proc. of the 35th Annual ACM Symposium on Applied Computing*, pages 2066–2073, 2020. 3

[5] Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote Photoplethysmograph Signal Measurement from Facial Videos Using Spatio-Temporal Networks. In *BMVC*, 2019. 1